DOCUMENT RESUME

ED 183 567 TM 010 013

AUTHOR Yen, Wendy M.
TITLE The Extent, Causes, and Importance of Context Effects on Item Parameters for Two Latent-Trait Models.
INSTITUTION CTB / McGraw-Hill, Monterey, Calif.
PUB DATE Nov 79
NOTE 43p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Context Clues; *Elementary School Mathematics; Grade 4; Grade 6; Intermediate Grades; *Item Analysis; Item Banks; Mathematical Models; *Reading Comprehension; Reading Tests; Statistical Analysis; Student Testing; Test Items
IDENTIFIERS *California Achievement Tests; *Latent Trait Models; One Parameter Model; Rasch Model; Test Analysis; Three Parameter Model

ABSTRACT

The extent, causes, and importance of context effects on item parameters for one- and three-parameter latent-trait models were examined. Items were taken from the California Achievement Tests Reading Comprehension and Mathematics Concepts and Applications subtests. The reading items were administered to 1,678 fourth-grade students, and the mathematics items were administered to 1,812 sixth-grade students. Context effects were measured by comparing the stabilities of item parameters obtained in the same context with the stabilities of parameters obtained in different contexts. Stability of item parameters was determined by testing the students twice with the same items. Changes in context were produced by changing the order of the items for the second test administration. Changes in context usually had substantial effects on the item parameters of both models, and to some extent these context effects appeared to be the result of the items' positions in the test booklets. Context effects appeared more important in item calibration than in item selection and in making predictions for single items than in making predictions or obtaining trait estimates with groups of items. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183567

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Extent, Causes, and Importance of Context Effects on Item Parameters for Two Latent-Trait models

Wendy M. Yen
CTB/McGraw-Hill

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Wendy M. Yen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running head: Context Effects on Latent-Trait Models

Manuscript submitted for publication: November 1979

TM010 013

Abstract

This paper examines the extent, causes, and importance of context effects on item parameters for one- and three-parameter latent-trait models. Items were taken from the California Achievement Tests (1977) Reading Comprehension and Mathematics Concepts and Applications subtests. The Reading items were administered to 1678 fourth-grade students, and the Math items were administered to 1812 sixth-grade students. Context effects were measured by comparing the stabilities of item parameters obtained in the same context with the stabilities of parameters obtained in different contexts. Changes in context usually had substantial effects on the item parameters of both models, and to some extent these context effects appeared to be the result of the items' positions in the test booklets. Context effects appeared more important (a) in item calibration than in item selection and (b) in making predictions for single items than in making predictions or obtaining trait estimates with groups of items.

There is widespread interest in latent-trait models---an entire edition of this journal (Summer, 1977) was devoted to them. One of the most important assumptions of latent-trait models is that the characteristics of each item can be described by one set of parameters. If this assumption is true, there is local independence of items. (See Allen & Yen, 1979, for a further discussion of the models and their assumptions.) If items are locally independent, they can be administered in any order without affecting performance on them, and items can be added to or taken from a test booklet without affecting the characteristics of other items. If examinees' performance on items, and, thus, the item parameters, are influenced by the sequencing of the items or by the characteristics of other items in the test, then context effects are occurring and local independence does not hold. It is the threefold purpose of this study to gauge the extent of context effects on latent-trait model item parameters, explore their cause(s), and evaluate their importance.

In this study context effects are measured by testing examinees twice with the same items. For some examinees the items are administered the same way both times; this is the "same context" condition. For other examinees the items are administered in different contexts on the two occasions, which produces the "different context" condition. For each condition the first- and second-testing item parameters are compared. If the item parameters are more similar in the same context condition than the different context condition, context effects have occurred.

There are several possible methods of comparing first- and second-testing item parameters, including correlations, square roots of mean squared differences, and significance tests. Each method offers a different perspective on context effects, but none of them gives a full view. A correlation reveals the strength of the linear relationships between two sets of item parameters, but a correlation cannot determine whether two sets of parameters are equal or whether they are unequal but linearly related. To examine differences in parameter values we can find the square root of mean squared differences (RMSD) of parameter values. Although a RMSD describes the size of the differences between parameters, it does not reveal whether the parameters are linearly related or whether differences between them are statistically significant. A significance test that takes into account the standard errors of the parameter estimates is informative, but such a test does not indicate the practical importance of any differences that are found.

In deciding how important differences in parameter values or context effects are, we need to have in mind the situation in which the parameter values will be used. Two situations in which context effects are particularly important are (a) item tryout and selection and (b) the calibration of items (i.e., the estimation of item parameters) for item pools. In an item tryout data are collected for items in a tryout booklet. Based on the tryout results, the best items are selected and placed in a final test booklet. The best items are those that are most discriminating (for the two- and three-parameter latent-trait models), have appropriate difficulty levels, and best fit the latent-trait model. If there are context effects, the items that appear to

be best in the tryout booklet may not be the best in the final booklet. If the items are recalibrated in the final booklet and these recalibrated item parameters are used whenever the final booklet is administered, context effects will not influence the scoring of the final booklet. In such a situation the only damage that context effects will do is in the possible selection of nonoptimal items.

Context effects can cause more problems with the calibration of items for item pools than with item tryouts. Item pools usually contain more items than the number feasible to administer to one group of examinees. Items for an item pool are calibrated by administering different sets of items to different groups of examinees. All the item parameters are linked to the same scale through a common subset of items called anchor or linking items. In the construction of an item pool the selection of the best items can also occur. After an item pool is created different examinees can take different items from the pool. With item pools, context effects can influence the choice of the best items and the paramter values associated with each item. The item parameter values influence the trait values that are obtained for any given examinee's pass/fail responses to the items, and the parameter values also influence the standard error of the trait value provided by the latent-trait model.

The item parameter values also influence the item-characteristic function (ICF) predicted by the latent-trait model. The values of the ICF can be important. For example, the ICF can be used to estimate the amount of improvement in a trait value that is needed

for the examinee to be likely to pass an important item. Or the ICF can be used to estimate the degree to which a rise in a trait value, say through average growth from grade 3 to grade 4, will affect an examinee's performance on particular items. In some cases an examinee's trait value is estimated on the basis of a subset of the items in an item pool. Using this trait estimate and the parameter values of the items the examinee did not take, predictions can be made of the probability the examinee would pass items in the pool that he or she did not take. Context effects can decrease the accuracy of all these predictions made on the basis of ICFs.

Test-characteristic functions (TCF) also can be influenced by context effects on item parameters. The TCF is the average ICF for the items in a particular test, and it relates proportion correct scores to trait values. The TCF is useful in interpreting raw score distributions (Lord & Novick, 1968, pp. 387-392). Context effects on item parameters will decrease the accuracy of interpretations based on the TCF.

In examining the extent, causes, and importance of context effects it may be important which latent-trait model is used. This study examines context effects for the three-parameter and one-parameter (or Rasch) logistic latent-trait models.

Method

Construction of Test Booklets

Items were chosen from the California Achievement Tests (1977) Forms C and D, Level 14 Reading Comprehension (Reading--80 items) and Level 16 Mathematics Concepts and Applications (Math--90 items). The Reading items all had four answer choices and the Math items all had five answer choices. Preliminary analyses were performed on data for students

who took both test forms (N=294 for Reading, N=379 for Math). Goodness-of-fit statistics were used to evaluate the items in terms of their fit to a two-parameter logisitic model; item difficulties and discriminations were reviewed to identify items with extreme difficulties or discriminations. Using these results, five sets of items were created in each content area. The first set of items (set A) had a range of difficulties and relatively good model fit, and these items would be used as anchor items for linking item parameters obtained in different booklets. The second and third sets of items (sets V and W) had relatively poor fit and, in some cases, extreme difficulties or low discriminations; these items would be included to alter item contexts. The fourth and fifth sets of items (sets X and Y) had relatively good model fit and discrimination, and nonextreme difficulties, and would be the items of major interest. Table 1 contains the numbers of items chosen for each set for each content area.

Insert Table 1 about here

Using these sets of items, fourteen booklets (seven in each content area) were created as described in Table 2. Within each

---

Insert Table 2 about here

---

content area the items in the different sets were intermingled within

the booklets, and the sequences of items were varied over booklets. Because of the connection of Reading items to passages and the connection of some Math items to graphs, there was necessarily some similarity over booklets in the local contexts for some items. The sequence of answer choices for an item was held constant over all booklets.

To indicate the degree of similarity of the sequence of X and Y items across booklets, the sequential positions of the X items and of the Y items were determined. Spearman rank-order correlations between items' positions in Booklet 1 and their positions in the other booklets are contained in Table 3.

Insert Table 3 about here

Test Administration

Students were tested in Grade 4 for Reading and Grade 6 for Math. Time limits for test administration were adjusted for the length of the booklet and were made comparable (on a time-per-item basis) to those given in the California Achievement Tests Examiner's Manual, Levels 14-19, Forms C & D (1977). For the first testing, each student took one of the seven booklets. Booklets 2 & 3 (as well as Booklets 4 & 5 and Booklets 6 & 7) were administered to students in the same classrooms on an alternate-seat basis. (This alternate-seat testing was done for a study of equipercentile equating that will not be reported here.) Two weeks later, all students took Booklet 1. The numbers of examinees with usable answer sheets for both first and second testings appear in Table 4.

Insert Table 4 about here

Models and Their Estimation

Two latent-trait models were used, the three-parameter logistic model and the one-parameter (Rasch) logistic model. For the three-parameter logistic model, the item-characteristic function (ICF) for the i-th item and k-th examinee is

$$P_i(\theta_k) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_k - b_i)}},$$

where θ_k is the latent-trait value for examinee k and a_i, b_i, and c_i are the discriminating power, difficulty, and lower asymptote, respectively, for item i. The ICF for the one-parameter model is

$$P_i(\theta_k) = \frac{1}{1 + e^{-1.7a\ (\theta_k - b_i)}},$$

where a is the item discriminating power common to all the items. To obtain latent-trait and item-parameter estimates, examinees' item-response vectors were analyzed using the LOGIST computer program provided by Wood, Wingersky, and Lord (1976). Trait estimates were not obtained for examinees who had zero or perfect scores, and using a default option of the Wood et al. program, trait estimates also were not obtained for examinees who did not answer at least a third of the items being scaled.

Linking Item Parameters

The X and Y item parameters were placed on the same scale using different procedures for Booklet 1 and Booklets 2 to 7. For Booklet 1, examinees took both X and Y items, and item-response vectors for the two sets were analyzed together, and their item parameters were automatically placed on the same scale. This procedure was followed when Booklet 1 was administered in the first testing or the second testing. In some analyses those examinees who took Booklet 1 on the first testing were

divided into two approximately equal-sized groups, and item parameters were obtained separately in the two groups. For convenience, these groups are labeled "Booklets 1A and 1B," even though the 1A and 1B test booklets were the same, and the "A" and "B" merely indicate different samples of examinees.

For Booklets 2 to 7, booklets were linked in pairs: 2 & 3, 4 & 5, 6 & 7. Booklets 2 and 3 can be used in an example of the linkage procedure. The responses to all the items in Booklets 2 & 3 were pooled and analyzed jointly. This was done by treating the items in the A, V, W, X, and Y sets as if they were all contained in one test booklet. An examinee's responses were used for all the items the examinee took, and the examinee was given a "not reach" for every item the examinee did not take. Examinees who took Booklet 2 were given "not reach" for items in sets W and Y, and those who took Booklet 3 were given "not reach" for items in sets V and X. Using the LOGIST program, an examinee's trait value was based only on the items the examinee actually took; "not reached" items were ignored. Similarly, an item's parameters were estimated using only the responses of examinees who "reached" the item. This joint analysis of Booklets 2 & 3 placed the parameters for the A, V, W, X, and Y items on the same scale.

For some analyses, responses for Booklets 2 to 7 were pooled and jointly analyzed. Thus, responses to Booklets 2, 4, and 6 entered into the estimation of the X item parameters, responses to Booklets 3, 5, and 7 entered into the estimation of the Y item parameters, and responses to all six booklets entered into the estimation of the A item parameters.

In summary, there were seven groups of first-testing item parameters: 1A; 1B; 1; 2 & 3; 4 & 5; 6 & 7; 2, 4, 6 & 3, 5, 7. All the second-testing data, which involved Booklet 1, was pooled to obtain one set of second-testing item parameters. Those examinees who took Booklet 1 (or 1A or 1B) on the first testing had the same context for the second testing; those who took any of Booklets 2 to 7 on the first testing had a different context for the second testing.

Comparisons Based on Item Parameters

Several types of comparisons between first- and second-testing item parameters were made: correlations, RMSDs, and chi-square tests of differences. ICFs and TCFs based on the first- and second-testing item parameters also were compared. Because correlations are unaffected by linear transformations of parameters, first- and second-testing parameters could be correlated without concern about whether the two sets of parameters were on the same scale. The other comparisons, such as RMSDs, were dependent on the actual parameter values. Thus, to make such comparisons it was necessary to place the parameters estimated from different samples and booklets on the same scale. To do this, a scale was arbitrarily chosen on the basis of the second-testing data. The second-testing trait values were transformed linearly so that the traits had a mean of zero and a standard deviation of one. The same linear transformation was made to the second-testing item parameters to make them consistent with the second-testing trait values. This procedure determined the scale, and it was then necessary to transform the first-testing parameters and traits so that they were on that same scale. To do this, the mean item difficulty and mean item discriminating power were obtained for the second-testing set

X items. Each of the seven groups of first-testing item parameters was linearly transformed so that its set X mean item difficulties and mean discriminating powers equaled the second-testing means. The first-testing trait estimates accompanying each group of first-testing item parameters also were linearly transformed to make them consistent with the item parameters. All the first-testing item parameters were then on the same scale as the second-testing parameters. Because the values of the set X parameters were used to place the first- and second-testing parameters on the same scale, it was not meaningful to then compare the set X parameter values for the first and second testings. However, it was meaningful to compare the set Y parameters for the first and second testings.

One comparison that was made was the root mean square difference (RMSD) between first- and second-testing item discriminating powers:

$$RMSD_{m,m'} = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{a}_{im} - \hat{a}_{im'})^2\right]^{\frac{1}{2}},$$

where a_{im} and $a_{im'}$ were item discriminating powers estimated for booklets m and m', and n items were compared. The RMSD also can be expressed as

$$RMSD_{m,m'} = \left[s_m^2 + s_{m'}^2 + (\bar{\hat{a}}_m - \bar{\hat{a}}_{m'})^2 - 2r_{mm'}\, s_m\, s_{m'}\right]^{\frac{1}{2}},$$

where s_m and $s_{m'}$ are the standard deviations of the item discriminating powers for the two booklets, $\bar{\hat{a}}_m$ and $\bar{\hat{a}}_{m'}$ are the means of the item discriminating powers for the two booklets, and $r_{mm'}$ is the corre-

lation between the item discriminating powers for the two booklets. RMSDs were determined in an analogous fashion for the item difficulties.

First- and second-testing item parameters were also compared taking into account the theoretical standard errors for the item parameter estimates (see Lord, 1979). The squared standard error of an estimated discriminating power is

$$SE^2(\hat{a}_i) \doteq \left\{ \left[\frac{1.7}{1 - c_i} \right]^2 \sum_{k=1}^{N} (\hat{\theta}_k - \hat{b}_i)^2 (\hat{p}_{ik} - \hat{c}_i)^2 \frac{(1 - \hat{p}_{ik})}{\hat{p}_{ik}} \right\}^{-1},$$

where $\hat{p}_{ik} = \hat{p}_i(\hat{\theta}_k)$ is the item-characteristic function evaluated using the estimated item parameters and trait values. The squared standard error of an estimated difficulty is

$$SE^2(\hat{b}_i) \doteq \left\{ \left[\frac{1.7a_i}{1 - c_i} \right]^2 \sum_{k=1}^{N} (\hat{p}_{ik} - \hat{c}_i)^2 \frac{(1 - \hat{p}_{ik})}{\hat{p}_{ik}} \right\}^{-1}.$$

The following chi-square statistic with n degrees of freedom was used to compare the first- and second-testing item discriminating powers for n items obtained in booklets m and m':

$$\chi^2_{m,m'} = \sum_{i=1}^{n} \frac{(\hat{a}_{im} - \hat{a}_{im'})^2}{SE^2(\hat{a}_{im}) + SE^2(\hat{a}_{im'})}.$$

Analogous chi-squares were obtained for the item difficulties. Note that the power of this chi-square test increases with increases in sample sizes.

The item-characteristic functions produced by first- and second-testing item parameter values were compared. The mean absolute difference between ICFs for booklets m and m' was

$$\frac{1}{k}\sum_{k=1}^{41}\left|\hat{p}_{im}(\theta_k) - \hat{p}_{im'}(\theta_k)\right|,$$

where θ_k took on values from -5 to +5 in steps of .25. Test-characteristic functions (TCFs) resulting from first- and second-testing parameter estimates also were compared. The mean absolute difference between test-characteristic functions for booklets m and m' was

$$\frac{1}{k}\sum_{k=1}^{41}\left|\hat{T}_m(\theta_k) - \hat{T}_{m'}(\theta_k)\right|,$$

where

$$\hat{T}_m(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_{im}(\theta_k).$$

To help understand the importance of context effects, standard error curves based on first- and second-testing item parameter values were compared.

$$SE(\hat{\theta}) = \left\{2.89\sum_{i=1}^{n}\frac{a_i^2(1-c_i)f_i(\hat{\theta})}{\left[1+f_i(\hat{\theta})\right]^2\left[1+c_i f_i(\hat{\theta})\right]}\right\}^{-1/2}$$

with $f_i(\hat{\theta}) = e^{-1.7a_i(\hat{\theta}-b_i)}$. These standard error curves are predicted from information theory and the latent-trait modes (Lord & Novick, 1968).

Results

Table 5 contains the numbers of examinees whose responses entered

---

Insert Table 5 about here

---

into the parameter estimations of the various booklets. These sample sizes usually are slightly smaller than those in Table 4 because examinees were not used if they did not answer at least a third of

the items or if they had zero or perfect scores.

Correlations between first- and second-testing X and Y parameter estimates are contained in Table 6[1]. An examination of plots

---

Insert Table 6 about here

---

associated with these correlations revealed that the item parameters were linearly related. Note that the correlations in Table 6 are a function of the particular items being correlated, and the correlations cannot be meaningfully compared across content areas. Correlations involving Booklets 1A, 1B, and 1 give information about the stability of parameters over a constant context. Correlations involving Booklets 2 to 7 give information about the stability of parameters over varying contexts.

Booklets 1A, 1B, 2 & 3, 4 & 5, and 6 & 7 all have item parameters estimated on the basis of about 200 examinees. For the discriminating powers the change in context produces a consistent drop in the correlations. The change in context also decreases the stability of the item difficulties for both the three-parameter and Rasch models. Examining the correlations for Booklets 2, 4, 6 & 3, 5, 7 we can see that when data are pooled over different contexts and sample sizes are thus tripled, there is an increase in the strength of the linear relationship between first- and second-testing item parameters. However, the correlations for Booklets 2, 4, 6 & 3, 5, 7 (which involve parameter estimates based on sample sizes of about 600) are

always lower than the corresponding correlations for Booklet 1 (which involve parameter estimates based on N≐400), and frequently lower than the correlations for Booklets 1A and 1B (N≐200). For Reading, the three-parameter and Rasch models have similar stabilities for item difficulties.

The first five rows of correlations in Table 6 deal with parameters estimated using the responses of about 200 examinees. For each column of the table we can compare the average squared correlation for parameters estimated in the same context (rows 1 and 2) with the average squared correlation for parameters estimated in different contexts (rows 3 to 5). This comparison can give us a rough estimate of the amount of variance in parameter estimates that is introduced by context effects. For the item discriminating powers the ratio of the average squared correlation for the different context condition to the average squared correlation for the same context condition ranged from .51 (Reading) to .66 (Math); the amount of variance shared between first- and second-testing item discriminating powers was much smaller for the different context condition than the same context condition. For the three-parameter item difficulties this ratio of average squared correlations ranged from .74 to .90, and for the Rasch difficulties this ratio ranged from .76 to .86. The context effects as measured by correlations appear to be about the same size for the Rasch and three-parameter difficulties; the context effects for the item discriminating powers are larger than the context effects for item difficulties.

The X and Y item parameters from Booklets 2 to 7 were linked by the use of the anchor items. It is possible that there were

inadequacies in the linking procedure that caused the reduced correlations between parameters estimated in these booklets and those in Booklet 1. A check on the importance of the linking procedure in affecting item parameter correlations was made. Correlations were found between the item parameters obtained for set X using the first-testing booklets and the parameters obtained for set X using the second-testing booklet, Booklet 1. Analogous correlations were obtained for set Y. There was no linking procedure influencing these item parameters. These correlations are contained in Table 7.

Insert Table 7 about here

In the vast majority of cases the correlations in Table 6 fall within the range of correlations in Table 7 produced by the X items and the Y items correlated separately. These results argue that the linking procedure did not cause the drop in item parameter correlations from Booklet 1 to Booklets 2 to 7.

RMSDs between set Y item parameters for the first and second testings are contained in Table 8. These results parallel those

---

Insert Table 8 about here

---

for the set Y correlations in Table 7, indicating the presence of context effects. Table 9 contains chi-square tests of the differences between the first- and second-testing set Y item parameters.

Insert Table 9 about here

The chi-squares theoretically have 20 degrees of freedom for Reading and 22 degrees of freedom for Math. Because chi-square values of 31.4 for Reading and 33.9 for Math are significant at the .05 level, the vast majority of the chi-squares in Table 9 are highly significant. The chi-square usually are much larger for the different context condition than the same context condition, and we can be more confident in rejecting the null hypothesis (i.e., that the parameters are the same) in the different context condition than in the same context condition.

For the three-parameter model different sets of item parameters can produce fairly similar ICFs. It is possible that the context effects that appear for the item parameters do not appear as context effects for ICFs.[2] It is also important to look for context effects for ICFs to help interpret the importance of context effects on item parameters. Mean absolute differences between first- and second-testing ICFs and between first- and second-testing TCFs are contained in Table 10. Context effects are apparent for the ICFs of both models.

Insert Table 10 about here

The TCFs show context effects for Reading but not for Math. The sizes of the mean absolute differences are smaller for the TCFs than the ICFs.

To obtain a better understanding of the sizes of the mean absolute differences in Table 10, an item was identified that had an ICF mean absolute difference of .036, which is approximately the median value for ICFs in Table 10. The first- and second-testing item-characteristic curves for this item are contained in Figure 1. Sim-

Insert Figure 1 about here

ilarly, a test booklet was identified that had a TCF mean absolute difference of .015. The first-testing test-characteristic curve for this booklet, and the second-testing test-characteristic curve appear in Figure 2. In Figure 3 are the predicted standard error

Insert Figures 2 and 3 about here

curves, SE(θ), for the tests appearing in Figure 2. The TCFs and SE(θ) curves are very similar for the first and second testings.

When the parameter values for a set of items change, trait estimates based on those items change. To obtain a clearer understanding of how much a change in parameter values alters trait estimates, trait estimates were obtained based on responses to the second-testing Reading Comprehension set Y items. Two sets of parameter values were used in these trait estimations, the first-testing Booklet 7 values and the second-testing Booklet 1 values. These two sets of parameter values were the least similar sets based

on the results in Tables 7 to 10. The correlations between the two sets of second-testing trait values were .93 for the three-parameter model and 1.00 for the Rasch model;[3] variations in the item parameter values had little effect on the relative values of examinees trait estimates. The means and standard deviations of the trait values were affected by the change in item parameters, and the RMSDs between the two sets of trait values were .58 for the three-parameter model and .56 for Rasch model.

Discussion

Recall the threefold purpose of this study: to gauge the extent, cause(s), and importance of context effects on latent-trait model item parameters. It was a consistent finding that item parameters estimated from the same context were more highly related than item parameters estimated from different contexts. These context effects were seen with correlations, root mean squared differences, and significance tests. The context effects appeared to be stronger for the item discriminating powers than the item difficulties, and they appeared to be equally strong for the three-parameter and Rasch item difficulties.

There are several factors that could have produced these context effects: (a) the inclusion of extraneous items (sets A, V, or W) in some of the booklets, (b) differences in the number of items scaled together, (c) systematic differences in the sequence in which items appeared in different booklets, and (d) other unspecified factors. Let us examine the evidence for and against the importance of these factors.

(a) Booklets 2 to 7 contained anchor items (set A items) that were not included in Booklet 1. It is possible that these items altered the trait measured by the booklet. For example, imagine that the Math set A items all were graph-reading items, and no graph-reading items appeared in sets X and Y. (This did not occur, but it gives an extreme example of how the set A items could alter the trait being measured.) On the face of it the set A items did not appear to have content systematically different from the X and Y items, but it is possible that they were statistically different from the X and Y items. If the set A items did alter the trait being measured, then the item parameters could have been affected. Booklets 2&3 contained not only set A items, but also sets V and W. Thus, Booklets 2&3 differed more in content from Booklet 1 than did Booklets 4&5 and 6&7. If the inclusion of extraneous items caused item parameters to change, then the relationship between Booklet 1 and Booklets 2&3 should be weaker than the relationship between Booklet 1 and Booklets 4&5 and 6&7. However, an examination of Tables 6 to 10 reveals that the relationship between Booklet 1 and Booklets 2&3 frequently are stronger than the relationships between Booklet 1 and Booklets 4&5 or 6&7. Thus, it does not appear that the inclusion of extraneous items in Booklets 2 to 7 was a major factor in reducing the strength of the relationships between item parameters for those booklets and Booklet 1.

(b) There were the same numbers of items calibrated in Booklets 1 and 2&3, but fewer items were calibrated in Booklets 4&5 and 6&7.

Item parameters may have been more stably estimated when more items were calibrated together. To examine this hypothesis the parameters for Booklets 2&3 were recalibrated excluding the V and W items. In this recalibration Booklets 2&3 had the same number of items as Booklets 4&5 and 6&7. The resulting parameters for Booklets 2&3 were correlated with second testing Booklet 1 parameters, and these correlations were compared with the corresponding correlations in Table 6. Only two of the six correlations changed: for Reading the three-parameter item discriminating power and difficulty correlations changed from .63 to .61 and from .76 to .77. Thus, it appears that number of items being calibrated did not have an important effect on the stability of the item parameter estimates.

(c) The sequence in which items appeared within a booklet could have had an influence on item parameter correlations. If sequence is important, we would expect item parameters obtained from booklets with similar item sequences to be more similar than item parameters obtained from booklets with dissimilar item sequences. Recall that Table 3 contains rank order correlations of item sequences between Booklet 1 and Booklets 2 to 7. These correlations produce expectations about which booklets should have parameters that are most highly related. The results in Tables 6 to 10 are partially consistent with these expectations. In particular the booklets with the most similar item sequences tend to have more strongly related item parameters than the booklets with the least similar item sequences.

These results indicate that the similarity of item arrangements might have an influence on the similarity of item parameters. Part of this influence could be the result of examinee fatigue or impatience to finish the test. If so, items should be relatively less difficult if they appear at the beginning of a booklet than at the end of a booklet. When estimating an item's parameters, LOGIST treats omitted items and not reached items differently. If an examinee omits, (does not answer) an item but does answer items that appear later in the booklet, the examinee is assumed to have a 1/A chance of having answered the item correctly, where A is the number of answer choices. If an examinee does not reach an item (i.e., he or she does not answer that item or any subsequent item), that examinee is not included in the calibration of that item. Thus, examinee fatigue or impatience can only affect an item's parameters if the examinee answers that item or subsequent items.

To examine the possibility of fatigue effects, a Reading passage was identified that appeared at the beginning of Booklet 1 and near the end of Booklet 7. The items for this passage were all relatively more difficult in Booklet 7 than Booklet 1. These items also had relatively higher discriminating powers in Booklet 7 than Booklet 1. It did not appear that speededness was an important factor because 93 percent of the examinees who answered at least a third of the items in Booklet 7 answered the last item in the booklet. A possible explanation is that a significant number of the examinees who answered questions about this passage near the end of Booklet 7 did not take

the care that examinees took when the passage was at the beginning of Booklet 1, and that for Booklet 7 items for this passage were important in discriminating between the higher scoring/more careful examinees and the lower scoring/less careful examinees. Notice that in Tables 7 to 9 Reading item parameters for Booklet 7 had particularly weak relationships with the parameters for Booklet 1.

Several other analyses like the one described in the previous paragraph were conducted. Items appearing at the end of booklets frequently but not always were relatively more difficult than the same items appearing at the beginning of another booklet. Results for item discriminations were not as systematic. Thus, it appeared that the location of an item in a booklet could have, but did not have to have, an impact on the item's parameters.

(d) Although the location of items in booklets appeared to be a partial explanation of context effects on item parameters, location did not appear to be a complete explanation. It is possible that other factors related to context could have influenced the parameters. For example, such factors might be specific to the particular content or difficulty level of preceding items. However, it is not apparent exactly what these factors are.

The importance of the context effects depends on the situation in which the item parameters are used. It appears that context effects will not cause serious problems with item tryouts in which the best items are selected but final item parameter values are not obtained. Even for tryout samples of about 200 examinees, the item difficulties

are stable enough so that items with desirable difficulties can be chosen. However, item discriminating powers are less stable than the item difficulties. If a test constructor wants to be confident of choosing the most discriminating items, tryout samples of more than 600 examinees should be used.[4] If it is not critically important that the most discriminating items must be chosen, but it is sufficient that the more discriminating items are likely to be identified, then tryout samples of 400 to 600 examinees are acceptable. Smaller tryout samples (say, N≐200) are acceptable for the Rasch model.

Context effects are more serious with item pools, where items are calibrated in booklets that are different from the booklets to be used subsequently. With such situations calibration sample sizes of at least 600 should be used. If item discriminating powers are to be estimated, calibration sample sizes of more than 600 should be used. Pooling calibration data over a variety of contexts should help reduce systematic context effects, such as the effects of fatigue. Increasing the calibration sample size in this manner will make item parameters more stable, but it appears to take a great increase in sample size to compensate completely for the item parameter instability produced by context effects.

Context effects tend to be more serious when dealing with one item than when dealing with a conglomeration of items. For example, predictions of performance on one item (e.g., Figure 1) tend to be more seriously affected by changes in item parameter values than

predictions of performance on a test (e.g., Figures 2 and 3). If a test user is particularly concerned about making predictions about individual items, it would be wise to use the same context for the calibration and later administrations of the item. If the same context is not used, then very large calibration samples, of more than 600 examinees, should be used.

Changes in item parameter values, such as those produced by context effects, tend to produce essentially linear transformations of trait estimates. These transformations affect the means and standard deviations of the trait values but do not seriously affect the relative sizes of examinees' trait values. If the scale produced for the trait values is important (for example, if the trait values are supposed to be equated to some other set of trait values), then it is advisable to use large item calibration samples and to maintain the same context for item parameter calibrations and trait estimates. If the trait scale is not important (for example, if a correlational study is being conducted), then smaller item calibration samples are acceptable and context effects on item parameters will not tend to have serious consequences.

Recall that one of the criteria in the selection of items for this study was fit of the items to the predictions of a two-parameter logistic model (in which $c_i = 0$ for all items). The V and W items were chosen from the items that had relatively poor fit to that model,

and the A, X, and Y items were chosen from the items that had relatively good fit. In theory an item that fits the two-parameter model will fit the three-parameter model, but will not necessarily fit the one-parameter model. This theory implies that the item selection procedure was biased in favor of the three-parameter model. In practice, however, this bias did not occur. Selection on the basis of fit to the two-parameter model had the effect of discarding items that had the worst fit for both one- and three-parameter models. The items that were retained for sets A, X, and Y were among those that had the best fit for both the one- and three-parameter models, but these items were not those that had systematically the best fit for either one of the models. In general the three-parameter model fit the data much better than the one-parameter model, so the items chosen for sets A, X, and Y fit the three-parameter model much better than the one-parameter model. Even if the items that had been chosen were those that had the best fit for the one-parameter model, the chosen items would have had the worse fit for the one-parameter than the three-parameter model.

Context effects were apparent for both the three-parameter and Rasch models. The item difficulties for the Rasch model tended to be more stable than item difficulties for the three-parameter model. If the Rasch model is used, smaller calibration sample sizes are needed. The ICFs and TCFs also tended to be more stable for the Rasch than the three-parameter model. This finding does not mean that the Rasch model was more accurate than the three-parameter

model. For example, consider an empirical item characteristic curve that changes slope from the first to the second testing. The Rasch model is constrained to predict a constant slope for the item for the two testings, while the three-parameter model will produce different slopes for the two testings. The three-parameter model ICFs will be less stable but more accurate (i.e. fit the data better) than the Rasch ICFs. This study presents information about context effects on the three-parameter and Rasch models, but other issues must also be considered when deciding which model is better for a given set of data (e.g., see Yen, 1979a, 1979b).

References

Allen, M. J., & Yen, W. M. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979.

California Achievement Tests, Forms C & D. Monterey, CA: CTB/McGraw-Hill, 1977.

California Achievement Tests Examiner's Manual, Levels 14-19, Forms C & D. Monterey, CA: CTB/McGraw-Hill, 1977.

Lord, F. M. Applications of item response theory to practical testing problems. New York: Lawrence Erlbaum Associates, Inc., 1979.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley, 1968.

Wood, R. L., Wingersky, M. S., & Lord, R. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton: Educational Testing Service, 1976.

Yen, W. M. The effects of context on latent-trait model item-parameter and trait estimates. Paper presented at the Computerized Adaptive Testing Conference, Minneapolis, Minnesota, June, 1979a.

Yen, W. M. Using simulation results when choosing a latent-trait model. Manuscript submitted for publication, 1979b.

Footnotes

[1]For the three-parameter model the lower asymptotes, c_i, had homogeneous values centered at .20 for Reading and .15 for Math.

[2]I am grateful to Michael Levine for pointing out this possibility.

[3]With the Rasch model the trait estimates are nonlinear transformations of number-right scores. Because the two sets of Rasch second-testing trait estimates were based on the same number-right scores, examinees' rank order positions had to be the same for the two trait estimates. However, the two sets of Rasch trait estimates could be nonlinearly related.

[4]In this study the effects of increasing sample sizes were examined by pooling data for different first-testing booklets and, therefore, different contexts. Recommendations about the sample sizes needed for various situations should be interpreted in light of the fact that we do not know what the effects would have been of increasing the sample size for one of the first-testing booklets or contexts.

Table 1

Numbers of Items in Each Set

Item Set	Numbers of Items	
	Reading	Math
A	10	11
V	10	11
W	10	11
X	20	22
Y	20	22

Table 2

Composition of Test Booklets

Booklet	Item Sets	Number of Items	
		Reading	Math
1	X+Y	40	44
2	A+X+V	40	44
3	A+Y+W	40	44
4	A+X	30	33
5	A+Y	30	33
6	A+X	30	33
7	A+Y	30	33

Note: Within each content area items from different sets are intermingled and presented in a different order in different booklets. There are seven Reading booklets and seven Math booklets.

Table 3

Spearman Rank-Order Correlations Between the Position of X or Y Items in Booklet 1 and the Position of Those Items in Booklets 2 to 7

Booklet	Item Set	Correlations with Booklet 1	
		Reading	Math
2	X	.90	.05
3	Y	.56	.40
4	X	.59	-.06
5	Y	.70	.62
6	X	.10	-.06
7	Y	.30	.03

Note. Each correlation is based on 20 items for Reading and 22 items for Math.

Table 4

Numbers of Examinees with Usable Answer Sheets for Both First and Second Testings

First-Testing Booklet[a]	Reading (Grade 4)	Math (Grade 6)
1	470	450
2	225	228
3 [b]	216	232
4	193	230
5 [b]	198	219
6	186	232
7 [b]	190	221

[a]All second testings used Booklet 1 and occurred two weeks after the first testing.

[b]These two Booklets were administered on an alternate-seat basis.

Table 5

Sample Sizes Involved in Parameter Estimations

Booklet	Sample Sizes	
	Reading	Math
First Testing		
1A	232	225
1B	228	224
2 & 3	223 & 214	228 & 232
4 & 5	187 & 195	228 & 218
6 & 7	183 & 184	228 & 218
1	460	449
2,4,6 & 3,5,7	593 & 593	684 & 668
Second Testing		
1	1660	1810

Note. Some of these sample sizes are slightly smaller than the corresponding sample sizes in Table 4 because examinees are excluded if they had zero or perfect scores or if they did not answer at least a third of the items.

Table 6

Correlations of First Testing X and Y Item Parameters with Second-Testing X and Y Item Parameters

First Testing		Correlations				
		Reading			Math	
N[a]	Booklet	Discriminating Power	Difficulty 3-par	Difficulty Rasch	Discriminating Power	Difficulty 3-par	Difficulty Rasch
200	Same Context 1A	.76	.95	.92	.76	.92	.97
	1B	.72	.84	.89	.78	.96	.99
200	Different Context 2&3	.63	.76	.82	.67	.91	.92
	4&5	.60	.81	.81	.59	.91	.92
	6&7	.30	.73	.73	.61	.84	.89
400	Same Context 1	.81	.92	.92	.80	.96	.99
600	Different Context 2,4,6 & 3,5,7	.67	.89	.90	.76	.95	.95

Note. Correlations are based on X and Y items, 40 items for Reading and 44 for Math. "Same Context" means the items were administered in the same context for the first and second testings, and "Different Context" means the items were administered in different contexts for the first and second testings.

[a]These are approximate sample sizes used to obtain first-testing item parameters; exact first-and second-testing sample sizes are contained in Table 5.

Table 7

Correlations of First Testing Item Parameters with Second-Testing Item Parameters

First Testing		Correlations					
		Reading			Math		
N[a]	Booklet	Discriminating Power	Difficulty 3-par	Difficulty Rasch	Discriminating Power	Difficulty 3-par	Difficulty Rasch
				X Items			
200	Same Context						
	1A	.70	.95	.92	.72	.90	.98
	1B	.75	.79	.89	.65	.97	.99
200	Different Context						
	2	.58	.90	.91	.67	.91	.93
	4	.60	.84	.85	.60	.93	.94
	6	.44	.83	.86	.67	.90	.94
400	Same Context						
	1	.79	.90	.92	.67	.96	.99
600	Different Context						
	2,4,6	.70	.95	.96	.81	.96	.96
				Y Items			
200	Same Context						
	1A	.85	.95	.94	.80	.96	.98
	1B	.69	.91	.91	.86	.95	.99
200	Different Context						
	3	.66	.78	.86	.76	.91	.91
	5	.55	.78	.76	.60	.91	.90
	7	.17	.63	.58	.61	.77	.81
400	Same Context						
	1	.85	.95	.94	.89	.97	.99
600	Different Context						
	3,5,7	.62	.85	.83	.78	.95	.93

Note. "Same Context" means the items were administered in the same context for the first and second testings, and "Different Context" means the items were administered in different contexts for the first and second testings.

[a]These are approximate sample sizes used to obtain first-testing item parameters; exact first-and second-testing sample sizes are contained in Table 5.

Table 8

Root Mean Squared Differences Between Set Y
Item Parameters for First and Second Testings

First Testing		Root Mean Squared Differences					
		Reading			Math		
N[a]	Booklet	Discriminating Power	Difficulty 3-par	Difficulty Rasch	Discriminating Power	Difficulty 3-par	Difficulty Rasch
200	Same Context						
	1A	.36	.23	.23	.26	.39	.23
	1B	.35	.32	.31	.31	.42	.15
200	Different Context						
	3	.34	.70	.46	.32	.35	.37
	5	.42	.50	.37	.38	.37	.42
	7	.64	.76	.63	.38	.65	.48
400	Same Context						
	1	.25	.22	.25	.23	.30	.18
600	Different Context						
	3,5,7	.32	.35	.30	.27	.27	.33

Note. "Same Context" means the items were administered in the same context for the first and second testings, and "Different Context" means the items were administered in different contexts for the first and second testings.

[a]These are approximate sample sizes used to obtain first-testing item parameters; exact first-and second-testing sample sizes are contained in Table 5.

Table 9

Chi-Square Tests of Differences Between Set Y Item Parameters for First and Second Testings

First Testing		Chi-Squares					
		Reading			Math		
N[a]	Booklet	Discriminating Power	Difficulty 3-par	Difficulty Rasch	Discriminating Power	Difficulty 3-par	Difficulty Rasch
200	Same Context						
	1A	37	25	54	46	69	49
	1B	38	77	94	106	40	22
200	Different Context						
	3	120	212	209	90	94	126
	5	151	118	133	95	108	157
	7	85	986	391	158	187	203
400	Same Context						
	1	27	56	113	60	70	56
600	Different Context						
	3,5,7	120	205	194	129	130	201

Note. "Same Context" means the items were administered in the same context for the first and second testings, and "Different Context" means the items were administered in different contexts for the first and second testings.

[a]These are approximate sample sizes used to obtain first-testing item parameters; exact first-and second-testing sample sizes are contained in Table 5.

Table 10

Mean Absolute Differences Between Set Y First- and Second-Testing Item-Characteristic Functions (ICF) and Test-Characteristic Functions (TCF)

First Testing		Mean Absolute Differences							
		Reading				Math			
		ICF		TCF		ICF		TCF	
N[a]	Booklet	3-par	Rasch	3-par	Rasch	3-par	Rasch	3-par	Rasch
200	Same Context								
	1A	.024	.021	.007	.010	.036	.022	.022	.014
	1B	.030	.030	.015	.016	.034	.015	.020	.010
200	Different Context								
	3	.054	.048	.040	.043	.041	.033	.020	.011
	5	.053	.037	.029	.011	.042	.037	.016	.027
	7	.066	.054	.051	.048	.058	.045	.029	.010
400	Same Context								
	1	.020	.023	.008	.012	.025	.019	.017	.013
600	Different Context								
	3,5,7	.040	.030	.011	.004	.030	.027	.005	.007

Note. "Same Context" means the items were administered in the same context for the first and second testings, and "Different Context" means the items were administered in different contexts for the first and second testings.

[a]These are approximate sample sizes used to obtain first-testing item parameters; exact first-and second-testing sample sizes are contained in Table 5.

Figure Captions

Figure 1. Item characteristic curves for an item with a mean absolute difference of .036 between first- and second-testing ICFs.

Figure 2. Test characteristic curves for a booklet with a mean absolute difference of .015 between first- and second-testing TCFs.

Figure 3. Standard error curves for the test booklets whose test characteristic curves appear in Figure 2.

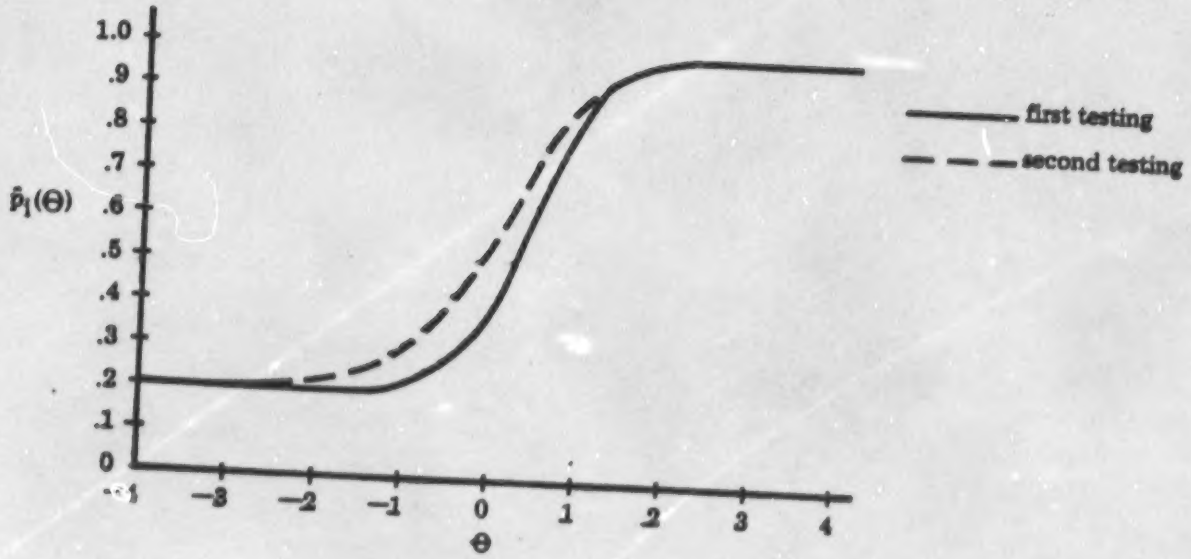
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
$\hat{p}_i(\Theta)$
−3
−2
−1
0
1
2
3
4
Θ
first testing
second testing

Fig. 41

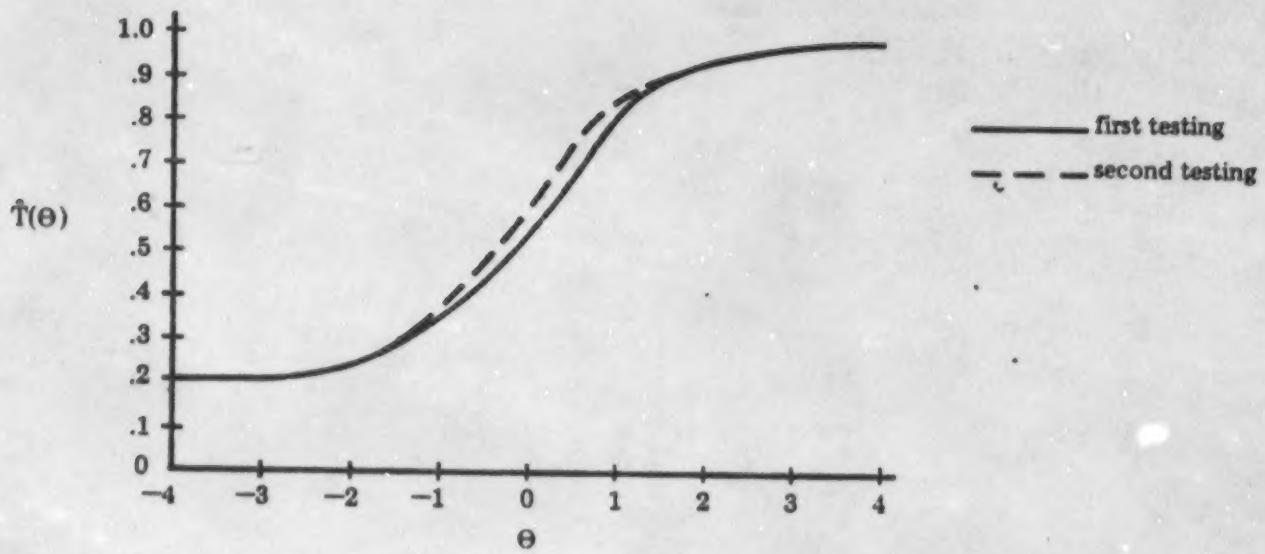

Fig. 2

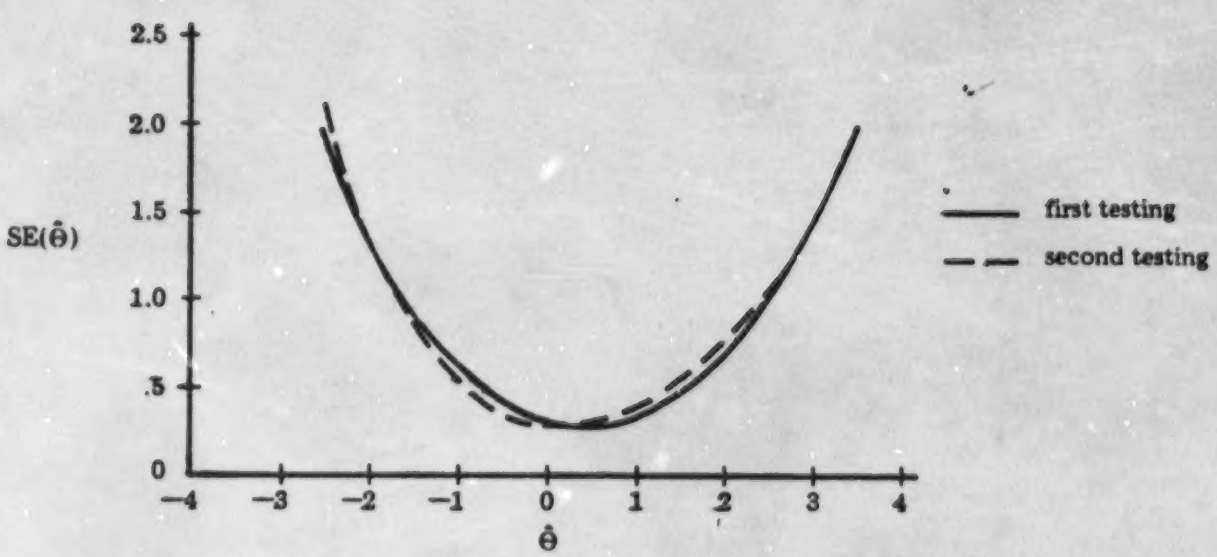

Fig 3